Multiplexing mechanical and translational cues on genes

Zuiddam, M.; Shakiba, B.; Schiessel, H.

Article

Multiplexing mechanical and translational cues on genes

Martijn Zuiddam,[1] Bahareh Shakiba,[1] and Helmut Schiessel[2,*]
[1]Institute Lorentz for Theoretical Physics, Leiden University, Leiden, the Netherlands and [2]Cluster of Excellence Physics of Life, TU Dresden, Dresden, Germany

ABSTRACT   The genetic code gives precise instructions on how to translate codons into amino acids. Due to the degeneracy of the genetic code—18 out of 20 amino acids are encoded for by more than one codon—more information can be stored in a basepair sequence. Indeed, various types of additional information have been discussed in the literature, e.g., the positioning of nucleosomes along eukaryotic genomes and the modulation of the translating efficiency in ribosomes to influence cotranslational protein folding. The purpose of this study is to show that it is indeed possible to carry more than one additional layer of information on top of a gene. In particular, we show how much translation efficiency and nucleosome positioning can be adjusted simultaneously without changing the encoded protein. We achieve this by mapping genes on weighted graphs that contain all synonymous genes, and then finding shortest paths through these graphs. This enables us, for example, to readjust the disrupted translational efficiency profile after a gene has been introduced from one organism (e.g., human) into another (e.g., yeast) without greatly changing the nucleosome landscape intrinsically encoded by the DNA molecule.

SIGNIFICANCE   DNA can contain several layers of information in addition to the classical genetic code. In this study we investigate two additional types of information that may exist on top of protein-coding information: mechanical information (in the form of nucleosome positioning) and translation efficiency. We ask to which extent these layers influence each other. Furthermore, we aim to restore the three layers of information of a protein-coding sequence when such a sequence is put in a host organism. To achieve this we use a graph-theoretical approach to manipulate genome sequences.

INTRODUCTION

As early as 1989 it was suggested by Edward N. Trifonov that DNA could carry several codes in addition to the classical genetic code (1). In particular, he mentioned a translation framing code (an excess of G in the first codon position), a chromatin code (caused by curved DNA) and a putative loop code (so as not to allow RNA secondary structure). In addition, overlapping genes were mentioned. Typically, however, the various scientific communities focus only on one additional layer of information. To give two examples: there exists a large body of work on DNA mechanics and geometry and how they influence the positioning of nucleosomes along DNA (mentioned in (1) as chromatin code) and another large body of work on the translational speed/efficiency in ribosomes and how it affects cotranslational folding. The question remains, howev-er, to which extent such different codes can really coexist on top of one another. This study answers this question using the examples of nucleosome positioning and translation efficiency.

The nucleosome is the repeated basic structure in chromatin. It is a stretch of DNA with a length of 147 basepairs (bp) wound 1 and 3/4 turns around a cylindrical aggregate made up of 8 histone proteins (2). The resulting disk-like complex is connected to the next such DNA spool by a short stretch of linker DNA. Notably, the wrapping length in the nucleosome is close to the DNA persistence length of about 150 bp or 50 nm. Bending a persistence length of DNA nearly two turns is quite expensive. Furthermore, the free energy of bending depends on the basepair sequence, which reflects the fact that the geometry and elasticity of the DNA double helix depend on sequence (3). This enormous sequence-dependent bending cost is compensated by the binding of the DNA molecule to the histone octamer at 14 binding sites (2). The binding is mainly to the DNA backbones, the chemistry of which is not dependent on the sequence. Taken together, this suggests that the affinity of

a given DNA sequence to be part of a nucleosome compared with another sequence is directly related to differences in the sequence-dependent bending costs. This makes it possible to write mechanical cues along DNA molecules to direct nucleosomes to occupy or to avoid certain positions. This has been referred to as the "nucleosome positioning code" (4) (for earlier versions of this idea, see e.g., (5) and (6), and for a review see (7)).

After reconstituting nucleosomes from DNA and histone proteins using salt dialysis, positional preferences of nucleosomes along genomic DNA can be clearly observed. By creating nucleosome maps using genome-wide assays that extract DNA stretches that were stably wrapped in nucleosomes (see e.g., (8)), one gets the nucleosome occupancy at each basepair position, which is the probability that the corresponding basepair is covered by a nucleosome. Two types of nucleosome positioning along DNA are found: rotational and translational positioning (9). Rotational positioning mainly reflects the fact that a given DNA stretch is typically not inherently straight because of the intrinsic geometries of the basepair steps involved. Nucleosomes therefore prefer positions where the DNA is prebent in the wrapping direction, resulting in sets of positions 10 bp (the DNA helical repeat) apart.

The specific basepair rules for rotational nucleosome positioning are typically formulated in terms of dinucleotides; rotationally positioned nucleosomes have an increased probability to feature GC steps (nucleotide G followed by nucleotide C) at positions where the major groove faces the protein cylinder (every 10th bp), and TT, AA, and TA where the minor groove faces the cylinder (4). A simulation of a nucleosome model that takes sequence-dependent DNA properties into account actually predicted these rules (10), and a simplified version of this nucleosome model made it possible to show analytically that these rules follow from the intrinsic shapes of the different basepair steps together with the fact that every basepair is part of a longer basepair sequence (11). In addition, we have explicitly shown that the degeneracy of the genetic code allows rotational positioning cues to be freely placed on top of genes without altering the resulting amino acid chain (12).

On the other hand, the translational positioning of nucleosomes is caused by DNA stretches that, overall, have a higher affinity for nucleosomes. It is known that this correlates well with their GC content (13–16). The physics behind the translational positioning is less clear than that of the rotational one; a recent study suggests that it is more about entropy than energy (17). There are various examples for translational mechanical cues, e.g., nucleosome-depleted regions before transcription start sites in unicellular organisms, which facilitate transcription initiation (8,16), mechanically encoded retention of a small fraction of nucleosomes in human sperm cells, which allows for the transmission of paternal epigenetic information (18),

and the positioning of six million nucleosomes around nucleosome-inhibiting barriers in human somatic cells (15).

Nucleosome positioning and stability are important not only at transcription start sites (and termination sites (8)), but also in between. It is known that elongating RNA polymerases slow down when they encounter nucleosomes. Instead of displacing them, polymerases get around nucleosomes through a looped intermediate structure (19–21). To enter the wrapped DNA inside the nucleosome, RNA polymerase waits for spontaneous opening fluctuations (a phenomenon called nucleosome breathing or site exposure (22)) and then rectifies these fluctuations (23). Nucleosome breathing is extremely sensitive to the basepair sequence of the wrapped DNA (24). Consequently, nucleosomes with highly asymmetric bending energies of their two wrapped DNA halves, such as Widom's famous 601 nucleosome (25), show highly asymmetric breathing from their two DNA ends (26,27) and, remarkably, act as polar barriers for RNA polymerases (28), since they only allow polymerases to transfer effectively in one direction.

In general, in eukaryotes, the GC content of exons is, on average, higher than that of introns (29), which in turn means that nucleosomes on exons are, on average, slower to cross for RNA polymerases than nucleosomes on introns. This has consequences for cotranscriptional events such as backtracking (allowing for error correction (30)) and alternative splicing (31,32). A comparative genomics study even concluded that local differences in nucleosome stability were amplified by GC content through evolution to establish new exons (33).

Important is also the fact that histone octamers can spontaneously change their positions along DNA, a phenomenon called nucleosome sliding (34). In this way, nucleosomes sample different positions, allowing for a rather slow equilibration of nucleosomes *in vitro*, at least locally (17). Two mechanisms have been suggested, both are based on thermally induced defects inside the nucleosome: single basepair twist defects (a missing or an extra basepair) (35–37) and 10 bp bulges (38,39). Recent simulation studies (40–42) found that both mechanisms can be at play and that it depends on the underlying basepair sequence which one is the preferred mechanism. Also a new experiment (43) indicates two types of movements of nucleosomes along DNA, small-scale repositioning on short timescales and longer ranged repositioning events on the timescale of minutes.

Importantly, *in vivo* there are chromatin remodelers present that use ATP to move nucleosomes along DNA. New experiments (44–46) and simulations (47) suggest that at least some of them induce twist defect pairs inside the nucleosome. Chromatin remodelers might help nucleosomes to equilibrate their locations along DNA (48), but they might also perturb the intrinsically preferred positioning of nucleosomes, together with other proteins that compete for DNA target sites (14). In addition, pioneer

transcription factors that can bind to nucleosomal DNA might play a role in recruiting remodelers (49).

In addition to classical genetic information and mechanical information, translation efficiency is encoded on the DNA. A gene on the DNA is transcribed and spliced such that it becomes a mRNA, which is then translated one codon at a time by the ribosomes. This creates an amino acid chain by facilitating the attachment of tRNAs containing the correct anticodon to the corresponding codons. The rate at which amino acids are attached to the growing amino acid chain is codon dependent and can be changed (over the course of evolution) since synonymous codons can have different attachment rates. This is because the translation efficiency of codons depends on the concentrations of corresponding tRNAs. These concentrations are correlated with the number of genes coding for the tRNAs (50). This is species specific, cell specific, and depends on the circumstances of the cells (51,52).

Translation efficiency has important consequences for the resulting proteins. Faster translation leads to larger amounts of protein, increased translational fidelity, less frameshifting, less amino acid misincorporation, less protein degradation, and less mRNA decay, while slower translation enhances cotranslational protein folding by giving more time for the protein to fold (52). Translation efficiency can affect the quality and quantity of proteins in many different ways. For instance: ribosome pausing can lead to ribosome collisions and cotranslational degradation of both mRNA and nascent chains (53). A number of protein functional and structural features are reflected in the patterns of ribosome occupancy, secondary structure, and tRNA availability along the mRNA (54). The same reference provides also specific examples where patterns of translation efficiency point to important structural and functional features of the corresponding proteins. An analysis of codon optimality in ten closely related yeasts has revealed universal patterns of conserved optimal and nonoptimal codons, often in clusters, which associate with the secondary structure of the translated polypeptides independent of the levels of expression (55). Further evidence that translation efficiency needs to be tuned comes from a study where the translation efficiency of a clock protein was optimized. While the protein levels increased, protein folding and function were affected (56). More examples can be found in recent reviews (57,58).

In our previous work (12) we studied the multiplexing of two layers, the layer of genetic information and the layer containing nucleosome positioning signals. However, due to the effects discussed above, we incorporate in this study also translation efficiency. This extra ingredient adds a new quality to the problem, which the other two layers (based on the genetic code and DNA mechanics) typically do not have: it is species specific. This leads to new questions, such as: Can a gene taken from a source organism be optimized for a host organism?

To explore how genetics, nucleosome positioning signals, and translation efficiency are multiplexed, we use graph representations of DNA sequences in combination with a shortest path algorithm. The methods in this work will be showcased using the gene tumor necrosis factor (TNF). First we discuss the multiplexing of genetics and nucleosome mechanics, as explored in previous work (12). Next we provide a short description of the translation efficiency model we use and how to use this model to obtain the highest and lowest possible translation efficiencies without changing protein-coding information. After this, we combine all three layers of information. We find how much the highest and lowest possible nucleosome energies on a gene are influenced by restrictions on translation speed. Finally we discuss genetically modified organisms. We change the DNA sequence of a gene, such that, when one puts this gene in a different organism, the genetic information is conserved while the mechanical information and translation efficiency landscape are close to their counterparts in the original organism. Also we change the sequence such that the translation efficiency landscape is extremely high or low, with mechanical information relatively unaltered and genetic information conserved.

METHODS

Model for nucleosome energy

To find out how genetics and nucleosome positioning signals are multiplexed, we revisit a method presented in previous work, where we showed how to obtain the lowest and highest possible nucleosome energy for a position on a gene without changing the resulting amino acid chain (12). We represented all possible sequences coding for the same amino acid chain as paths through a weighted directed graph in combination with a shortest path algorithm. The weights were given by a probabilistic trinucleotide model obtained through Monte Carlo simulations of a coarse-grained nucleosome model with sequence-dependent DNA elasticity (59), although any short-range probability or energy model may be used. In the trinucleotide nucleosome energy model (59) the energy cost of wrapping a sequence S of nucleotides $S_i \in \{A, T, C, G\}, i = 1, ..., L$ with $L = 147$ into a nucleosome is given by

$$E(S) = \sum_{n=1}^{L-2} E_n(S_n, S_{n+1}, S_{n+2}), \tag{1}$$

where the E_n values are energy costs associated with a trio of nucleotides, see (12) for details. The simulations (10) that generated the parameters of the trinucleotide model use a coarse-grained nucleosome model, where the wrapped DNA is represented by the rigid basepair model (3). The DNA is restricted by 28 constraints mimicking the binding of phosphates of the DNA backbone to the protein core. These constraints were extracted from a nucleosome crystal structure without adjustable parameters. In the rigid basepair model the conformation of the DNA molecule is described by the positions and orientations of the basepairs that are modeled as rigid bodies. This rigid basepair model assumes sequence-dependent nearest-neighbor interactions with energy costs quadratic in the deformations from the intrinsically preferred geometry (3). As in our previous work, we use a graph representation of all possible sequences that code for the same protein. To understand the new method we introduce in this work,
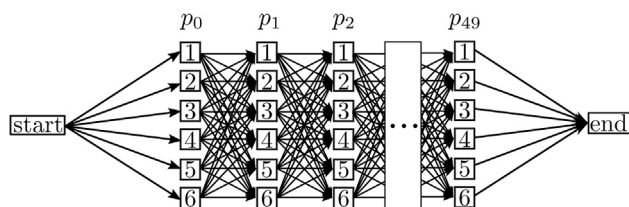
FIGURE 1 Graph $\mathscr{G}_E$ depicts all synonymous ways to encode a given amino acid sequence $p_0, p_1, ..., p_{49}$ wrapped around a histone core. Since each amino acid is encoded by three basepairs, and a nucleosome is of length 147, either 49 or 50 codons (together with the histones) form a nucleosome. For each amino acid six options are shown, representing the, at most, six possible ways to code for the same amino acid. The actual bases depend on the amino acid in question. When there are less than six options, one can simply leave out the surplus of nodes. Weights are assigned such that each path from *start* to *end* has a length equal to the total wrapping energy of the corresponding codon sequence.

we first briefly summarize the method we used before. The notation we use here is quite different, such that we can more easily incorporate translation speed in a graph.

The DNA on a nucleosome consists of 147 bp, which corresponds to either 49 or 50 codons. Suppose we have a sequence of 50 codons. These codons encode a sequence of amino acids $p_0, p_1, p_2, ..., p_{49}$. The number of different codons coding for the same amino acid is 6 at most. Therefore, the most general representation of all possible ways to code for the same protein at one nucleosome position is given by graph $\mathscr{G}_E$ in Fig. 1. In this figure, under each amino acid p_n, six numbers are shown representing the (at most) six possible codons, which we will refer to in the following as $p_n(1), p_n(2), ..., p_n(6)$. The actual basepairs of the codons depend on the amino acid in question. To obtain this graph we draw the following weighted edges: from start to $p_0(i)$ with weight zero for any i, from $p_{49}(i)$ to end with weight $w_{end}(p_{49}(i))$ for any i, and from $p_n(i)$ to $p_{n+1}(j)$ with weight $w_n(p_n(i), p_{n+1}(j))$ for any i, j and $n = 0, 1, ..., 48$. The weight w_i is given by

$$w_i(C, D) = E_{3i-2}(C_1, C_2, C_3) + E_{3i-1}(C_2, C_3, D_1) \\ + E_{3i}(C_3, D_1, D_2) \quad (2)$$

and the weight w_{end} by

$$w_{end}(D) = E_{145}(D_1, D_2, D_3) \quad (3)$$

where C_k and D_k denote the kth base of codons C and D. Now the length of a path from *start* to *end* in the graph equals the energy of a corresponding sequence. The lowest and highest energy can be found using a shortest path algorithm.

Model for translation efficiency

To add a single amino acid to the polypeptide chain, the ribosome goes through a cycle of chemomechanical reactions. A summary of distinct states and reversible/irreversible steps of the decoding and peptidyl transfer processes can be found in reviews (60,61). There exist many models for the speed or efficiency at which the polypeptide chain is created. One such model is the tRNA adaptation index (tAI) (62). The tAI is a measure for how well a gene is adapted to tRNA abundance in a cell. The abundances of different tRNAs are derived from the tRNA copy numbers, since it turns out that these quantities are correlated (50). The tAI also takes into account how efficient the wobble interactions between codons and tRNAs are, derived from gene expression in *Saccharomyces cerevisiae*. To improve upon this model, a species-specific tAI has been created, called stAI (63). As its name suggests, this method adjusts the tAI weights to any target model organism, not just yeast. Using stAI weights produces significantly better predictions for nonfungal protein abundance. In this work, we use weights W_C for the adaptiveness of codon C obtained from STADIUM: the Species-Specific tRNA Adaptive Index Compendium. In STADIUM, the codon values W_C have been precalculated for a huge range of species (64). We denote the translation efficiency/relative adaptiveness of a codon C by

$$T(C) = \frac{W_C}{W_{max}} \quad (4)$$

where W_{max} is the highest value for W of all codons (of all amino acids).

RESULTS AND DISCUSSION

Multiplexing of genetics and mechanics

In this work we study a gene from the human genome: the gene TNF, which codes for a cytokine. A cytokine is a signaling molecule involved in the immune response of mammals (65). TNF has an important role for both innate and adaptive immune responses, and is related to cancer progression and metastasis (66). TNF was chosen because
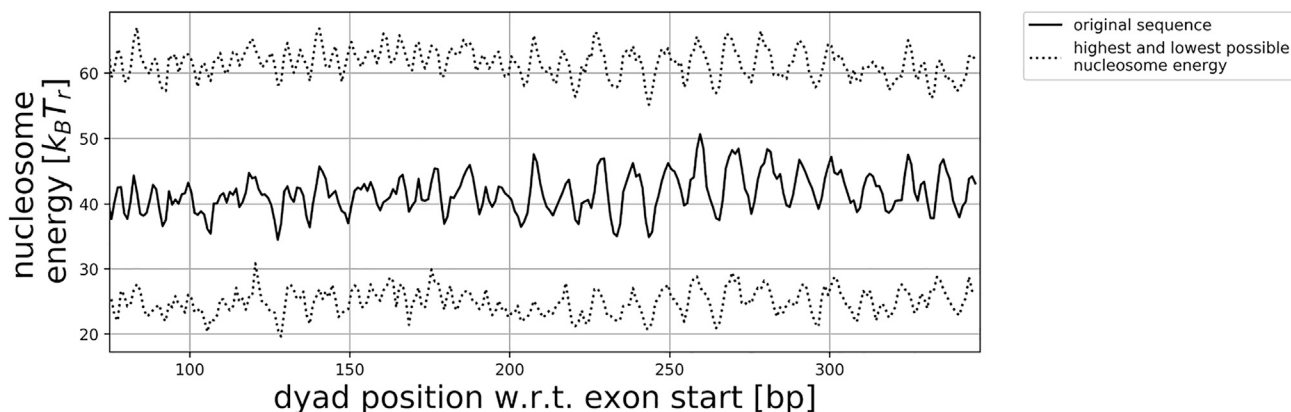

FIGURE 2 The nucleosome energy landscape for the fourth exon of the human gene tumor necrosis factor (TNF) is depicted by the solid line. The dotted lines give the highest and lowest possible energies at these positions for any theoretical exon coding for the same amino acid chain. The dyad position is the position of the central basepair on the nucleosome. Note that the actual energies lie roughly in the middle of their possible values, see text for details.
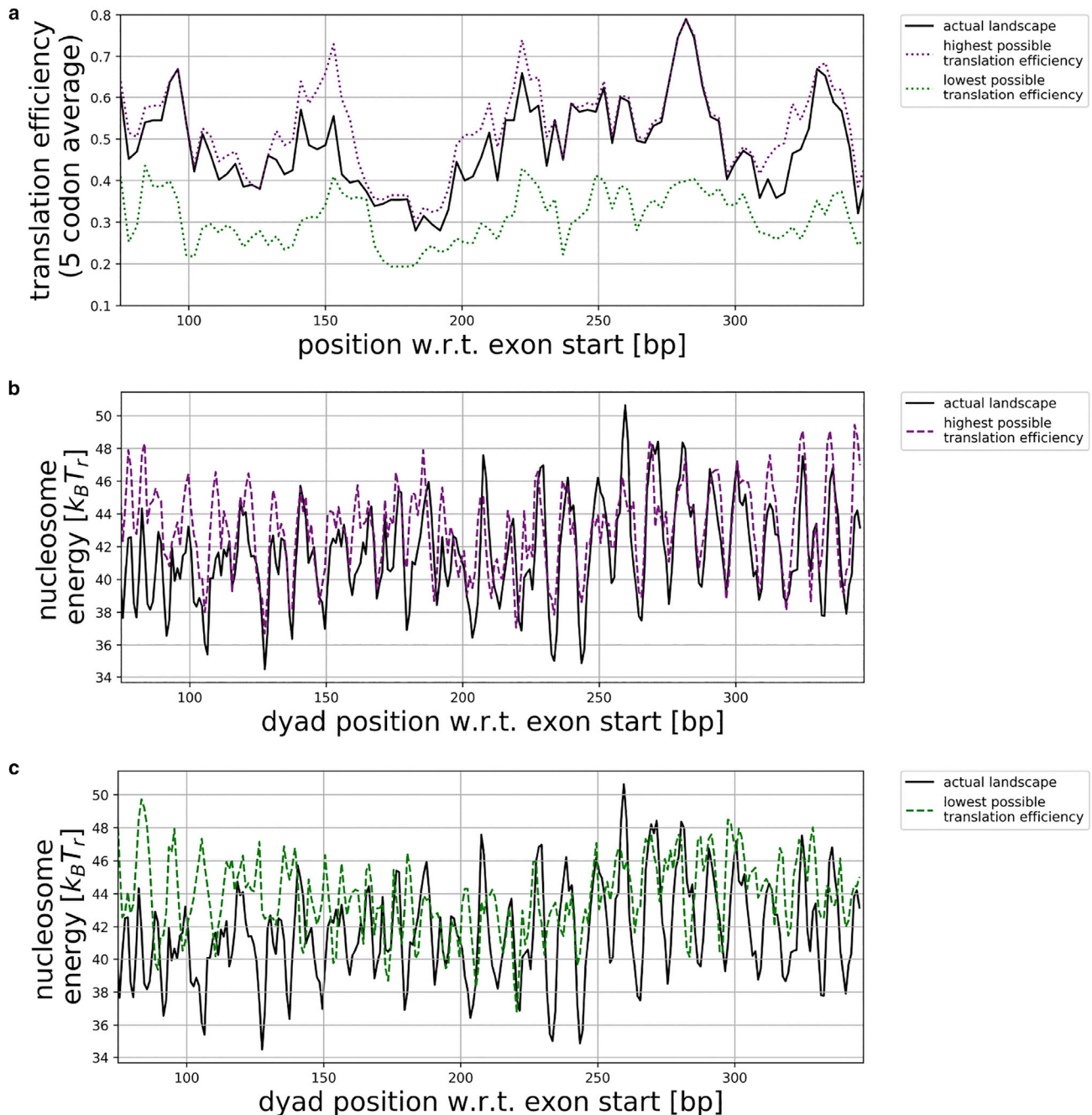
FIGURE 3 In (*a*), the translation efficiency landscape for the fourth exon of the human gene tumor necrosis factor (TNF) is depicted by the solid line. The dotted lines, in purple and green, denote, respectively, the highest and lowest possible translation efficiency when codons may be replaced by synonymous codons. We average over five codons to obtain a clearly visible signal. Changing the translation efficiency can have large consequences for the nucleosome energy landscape, as visible in (*b*) and (*c*). In (*b*) and (*c*), the solid line depicts the original nucleosome energy landscape. In (*b*), the (*purple*) dotted line depicts the energy landscape corresponding to the highest possible translation efficiency, (*c*) depicts (in *green*) the same but for the lowest efficiency. The energy landscape changes more when the translation efficiency is minimized. To see this figure in color, go online.

it is the second-most cited gene (67). The most cited gene, p53 (67), was not used because it has no exon significantly longer than the nucleosomal wrapping length. The fourth exon of TNF is much longer than the nucleosomal wrapping length, allowing us to safely ignore the effect of noncoding DNA on the nucleosome energy landscape.

Fig. 2 depicts the nucleosome energy landscape, calculated from Eq. 1, for the fourth exon of TNF. The dyad position is the position of the basepair in the middle of the nucleosome. The plot also depicts the highest and lowest possible energies at these positions for any theoretically possible exon, coding for the same amino acid chain. These
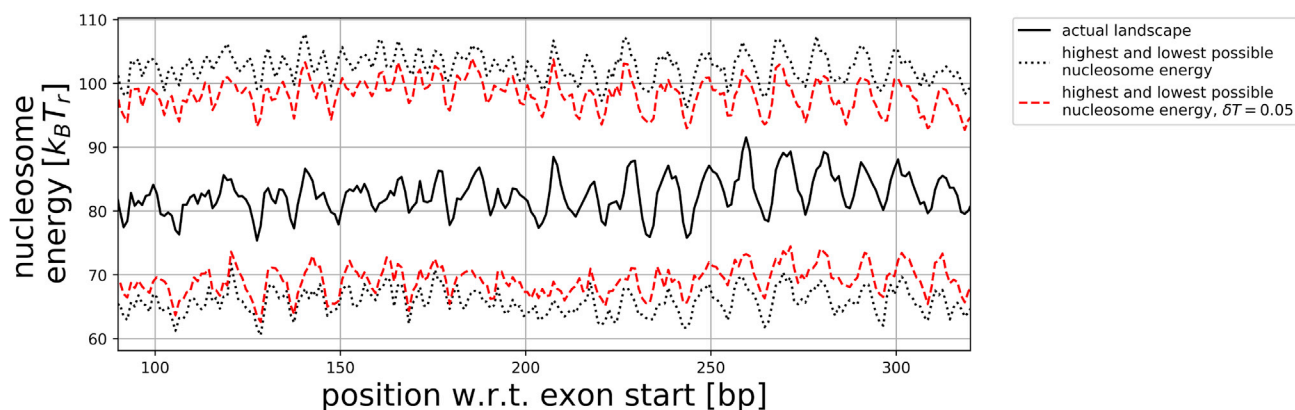
FIGURE 4 Same as Fig. 2, but with the addition of the highest and lowest possible nucleosome energy with a translation efficiency restriction of $\delta T = 0.05$. Note that there is still most of the range in nucleosomes energies available, even though the translation efficiency landscape is severely restricted. To see this figure in color, go online.

values were obtained by using a graph representation of all possible synonymous codons and a shortest path algorithm, following our method introduced in (12).

Note that all the nucleosome energies in the exon under investigation are somewhat around the middle of their possible values. This is consistent with what is known from experiments that compare the affinities of nucleosome positioning sequences. For example, the strong sea urchin 5S RNA gene nucleosome positioning sequence is still far from optimal. This can be seen by direct comparison with the well-known 601 sequence, an artificial sequence that has been selected from a large pool of random sequences for its capability to reconstitute nucleosomes (25). The 601 sequence has about 5 $k_B T_r$ net free energy gain relative to the 5S sequence (25). It is straightforward to improve the affinity of the 601 sequence further, e.g., by symmetrizing the highly asymmetric sequence (one half is much more strongly adsorbed than the other) or by adding just a few TA steps (27). But it is important to realize that optimizing sequences much more might make them less easy to reconstitute nucleosomes, which is probably the reason why they were not found in the experiment.

For similar reasons we believe that nucleosome energies are evolutionarily tuned, but that they are nowhere tuned for maximum or minimum stability. This provides the wiggle room for putting other layers of information on the DNA. Whereas the 10 bp periodicity in the nucleosome energy also occurs for random basepair sequences, larger-scale undulations of the mean or minimum values (as also seen for the exon under consideration) can point toward biologically meaningful signals. In the introduction, we gave some examples of how the resulting modulations in nucleosome stability might affect biological function.

Multiplexing of genetics and translation efficiency

The translation efficiency model we use does not contain dependencies on the neighbors of codons. Therefore, to obtain

the highest and lowest possible translation efficiency (keeping the protein intact) we can simply pick the codons with the highest and lowest efficiencies. The result for the fourth exon of TNF is depicted in Fig. 3 a. We average over five codons to obtain a clearly visible signal. It can be seen from the plot that the exon mostly favors a high translation efficiency. Remarkably, at some places the translation efficiency even reaches the maximum values, which is in contrast to the case of the nucleosome energies, which are nowhere near the extreme values (see Fig. 2).

When a sequence is altered to favor either high or low translation efficiencies, the nucleosome energy landscape is changed as well. In Fig. 3, b and c the new landscapes are shown together with the original landscape, the same as in Fig. 2. Since the translation efficiency is closer to being maximal, the energy landscape changes more when the translation efficiency is minimized (Fig. 3 c), compared with when it is maximized (Fig. 3 b). It is remarkable, however, that the nucleosome energy landscape, even in Fig. 3 c, deviates much less from the original landscape than what could be expected looking at the full range of possible energies in Fig. 2. A combination of two factors contributes to this: (1) the distribution of possible nucleosome energy values is sharply peaked and extreme values are only taken by a small fraction of sequences (see Fig. 2 b in (12)), and these sequences show more or less 10 bp periodic signals in, e.g., GC content (see Fig. 3 in (12)). (2) Translation efficiency contains no periodic signals and its optimization thus cannot produce the 10 bp periodic signals necessary for creating extreme nucleosome energies.

Multiplexing three layers of information: Genetics, mechanics, and translation efficiency

We now study the multiplexing of the three types of information. We have seen that the space of possible nucleosome energies for a gene is large. Now we investigate the very same while including the translation efficiency landscape.

FIGURE 5 (*a*) Translation efficiency landscape of the fourth exon of TNF in three organisms: the original (human) and two possible host organisms: yeast and *A. thaliana*. (*b*) The original landscape as well as the highest and lowest possible translation efficiency values in the hosts. We see that the original landscape cannot be reproduced in *A. thaliana* by looking at the highest and lowest values alone. To see this figure in color, go online.

What are the lowest and highest possible nucleosome energies when the translation efficiency landscape at any position may only change by no more than some fixed amount δT?

We calculate the energy cost of wrapping a codon sequence C around a nucleosome. A nucleosome of 147 bp corresponds to either 49 or 50 codons. We denote the codon sequence by $C = (C_0, C_1, ..., C_{49})$. We look at the set of sequences where the translation efficiency at any codon position (averaging over five codons) may only be altered by no more than some value δT:

$$\frac{1}{5} \left| \sum_{i=-2}^{i=2} T(C_{n+i}) - T(C_{n+i}^{\text{new}}) \right| \leq \delta T, \text{ for } n$$
$$= -2, -1, ..., 51, \qquad (5)$$

where C^{new} denotes any sequence of synonymous codons. We have included four neighboring codons on each side of the codon sequence, denoting them by C_i for $i < 0$, $i > 49$. (Including more codons does not make a difference for the results.)

Applying this restriction to a graph is not difficult. In previous work (12) we implicitly used the fact that genetic information can be considered as a restriction on the possible nodes of a graph: one can simply disallow nodes corresponding to nonsynonymous codons. We apply the same strategy for the translation efficiency: we disallow (or prune) nodes that do not conform to the efficiency restriction (see the supporting material). Again one can find the lowest and highest energy by calculating its shortest and longest paths. The result for TNF is depicted by Fig. 4. It shows that a strong restriction, $\delta T = 0.05$, results in only a small change in the highest and lowest possible energies. This shows that there is still a large wiggle room for changing nucleosome energies, even after severely restricting two other layers of information.

Genetically modified organisms

We have observed that changing the translation efficiency of a gene can have strong effects on its nucleosome energy landscape. On the other hand, we have also seen that there is quite some flexibility for the three layers of information—genetic information, mechanical information, and translation efficiency: restricting changes in translation efficiency would still allow a large range of possible nucleosome energy landscapes. We introduce now a method that uses this malleability in two scenarios with biological relevance. One scenario is the creation of highly efficient exons
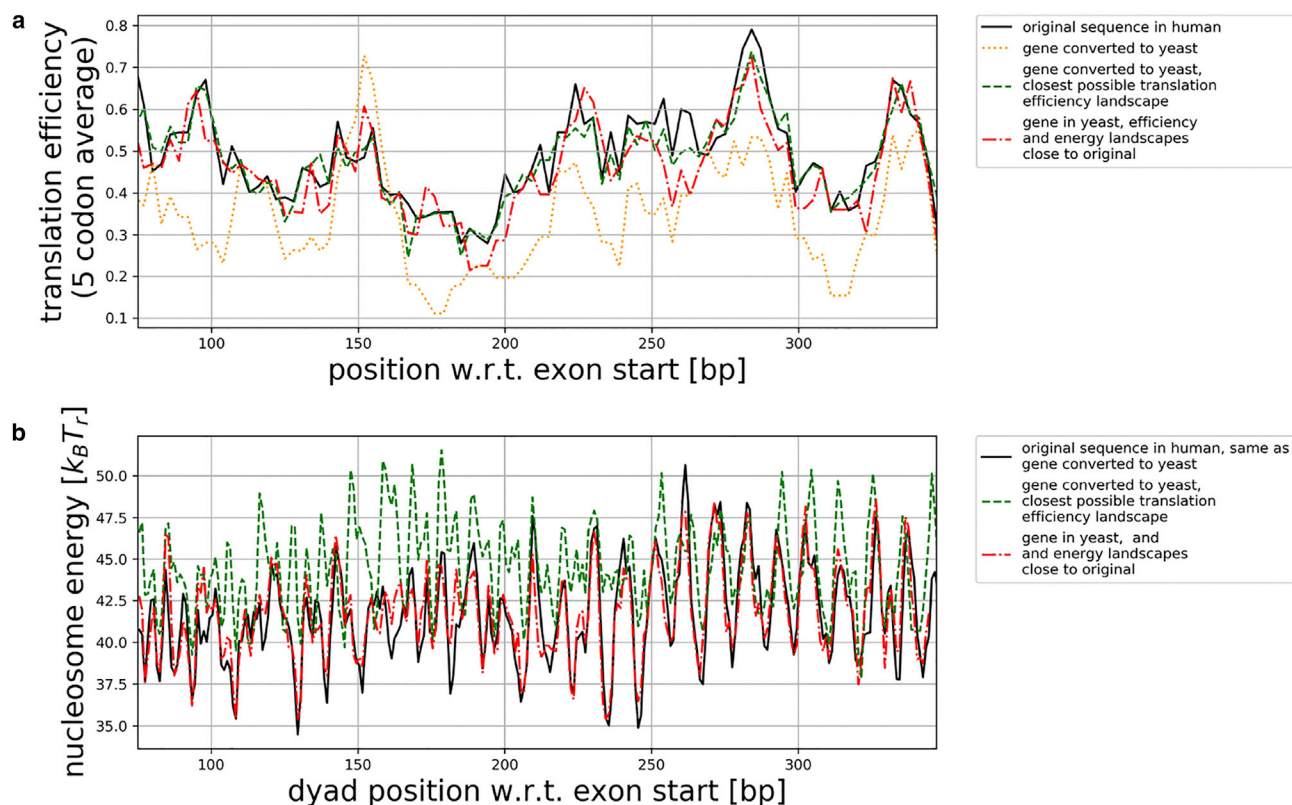
a

b

FIGURE 6 For the fourth exon of gene TNF, (*a*) depicts several translation efficiency landscapes and (*b*) the corresponding nucleosome energy landscapes. The original landscapes in human are depicted by a solid black line. The translation efficiency landscape of the original sequence in yeast is depicted by the orange dotted line. The closest possible translation efficiency landscape is depicted by the green dashed line. The corresponding nucleosome energy landscape is now quite different from the original landscape. A compromise is made for the red slash-dotted curves, where both landscapes closely resemble the original landscapes, using Eq. 10 with $c_T = 1$ and $c_E = 1/50000$ $[1/k_BT_r]$. To see this figure in color, go online.

while keeping the nucleosome energy landscape close to original. The other scenario is putting a gene in a different organism—a host organism—and making the three layers of information in the host close to how they were in the source organism. The method is best understood by studying the second scenario first.

Genes in host organisms

Since the conversion of codons to amino acids is practically universal, a gene in a host organism will almost certainly encode the same amino acid chain. Also, since the nucleosome energy landscape depends only on the physical properties of the basepair sequence, the nucleosome energy landscape, too, remains unchanged. However, the translation efficiency landscape, the third layer of information considered here, may be very different in a host organism. This is due to differences in tRNA concentrations between organisms. In Fig. 5 *a* we show that the shape of the translation efficiency landscape of TNF is qualitatively different in hosts *Saccharomyces cerevisiae* (baker's yeast) and *Arabidopsis thaliana* (thale cress, a plant and model organism). Our first goal is for the host organism to have all three layers

of information close to the original. More specifically, we want to make the translation efficiency landscape resemble the original landscape, without changing the amino acid sequence and while making only minor changes to the nucleosome energy landscape.

Translation efficiency in host organisms

Our first goal is to find out exactly how close the translation efficiency landscape in a host organism can get to the original landscape, ignoring for the moment the nucleosome energy landscape. It turns out that this can be a problem, as can be seen by inspecting the highest and lowest values of the translation efficiency for the gene TNF in host organisms in Fig. 5 *b*. In this figure we see that the original translation efficiency landscape fits almost everywhere inside the limits of host organism yeast. For the host *A. thaliana*, however, it is at many positions impossible to restore the translation efficiency of this gene without changing some of the amino acids.

We show now how close the translation efficiency landscape of yeast can get to the original while keeping amino acid information intact. We generalize by using the terms
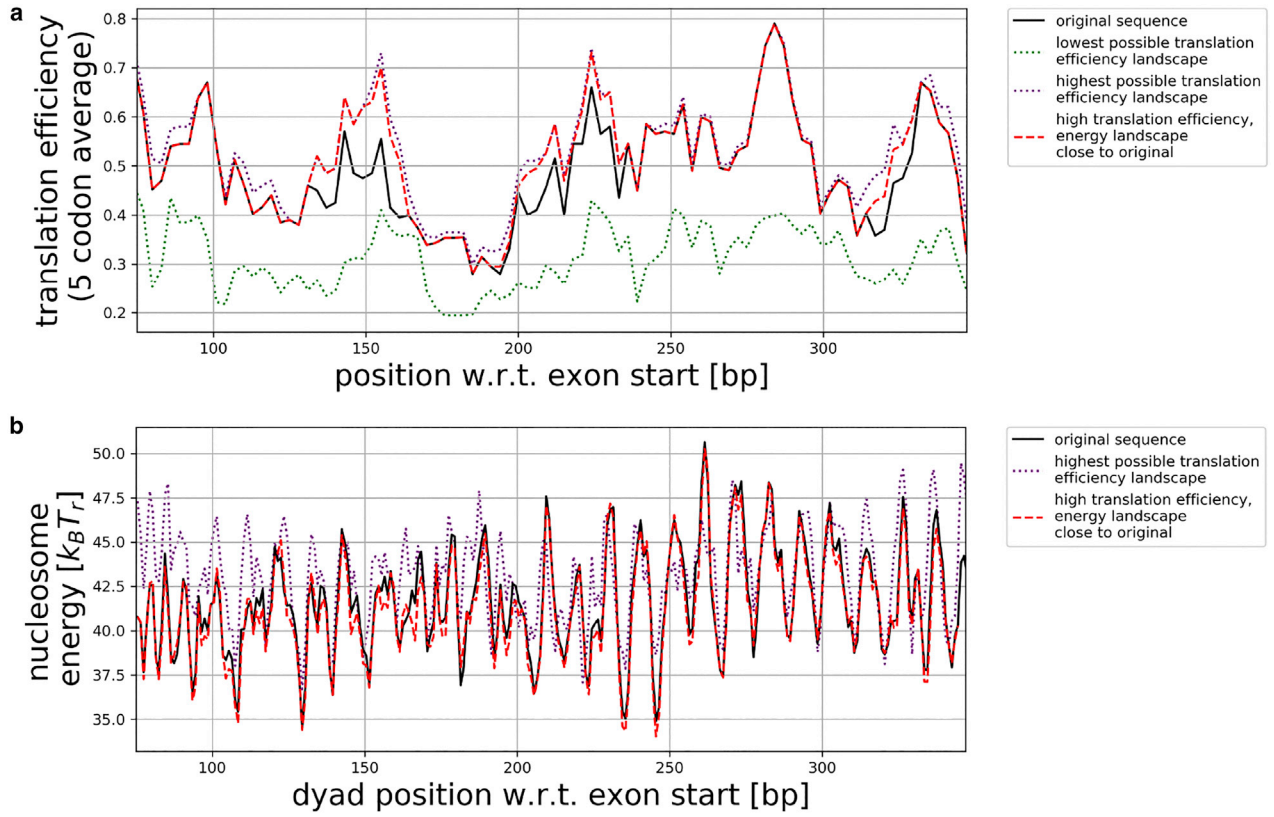
FIGURE 7 For the fourth exon of gene TNF, (*a*) depicts several translation efficiency landscapes and (*b*) the corresponding nucleosome energy landscapes. The original landscapes in human are depicted by a solid line. In (*a*), the dotted lines, in purple and green, denote, respectively, the highest and lowest possible translation efficiencies when codons may be replaced by synonymous codons. In (*b*), the nucleosome energy landscape as a result of the highest possible translation efficiency landscape is shown as a dotted purple line. The red interrupted lines in (*a*) and (*b*) show the result where the translation efficiency is high but the energy landscape is close to the original landscape, with the amino acid sequence intact. The red line is the result of using Eq. 13 with $c_T = 1$ and $c_E = 1/50000$ $[1/k_BT_r]$. To see this figure in color, go online.

host for yeast and *source* for human. Formally, we will minimize the distance D_T between the original translation efficiency landscape of a gene $G = (G_0, \ldots, G_{3N})$ in the source and the translation efficiency landscape of the gene G' in the host, where G' is a sequence that codes for the same amino acids as G. Here, N is the number of codons in G and G_i denotes the ith basepair.

Let A_G be the set of all sequences that code for the same amino acid chain as G. We choose the closest sequence G' such that

$$D_T(G, G') \leq D_T(G, X) \text{ for all } X \in A_G \quad (6)$$

with

$$D_T(G, X) \equiv \sum_{p=2}^{N-3} \Delta T_{\text{host}}^{\text{source}}(G, X, p), \quad (7)$$

where $\Delta T_{\text{host}}^{\text{source}}(G, X, p)$ describes the difference between the average translation efficiency of an altered sequence X in the

host and the original sequence G in the source, five codons centered around a codon position p:

$$\Delta T_{\text{host}}^{\text{source}}(G, X, p) \equiv \left| \sum_{i=-2}^{i=2} T_{\text{source}}\left(G_{3(p+i)}G_{3(p+i)+1}G_{3(p+i)+2}\right) \right.$$
$$\left. - T_{\text{host}}\left(X_{3(p+i)}X_{3(p+i)+1}X_{3(p+i)+2}\right) \right|. \quad (8)$$

Here, the subscript of T denotes for which organism the translation efficiency is calculated.

The resulting sequence G' corresponds to the translation efficiency landscape depicted by the green interrupted line in Fig. 6 *a* for TNF. The altered translation efficiency landscape in yeast is extremely close to the original landscape in human. As a side effect, however, such an optimized basepair sequence typically leads to dramatic changes in the nucleosome energy landscape, as can be seen for TNF by the green interrupted line in Fig. 6. For examples using other genes, see the supporting material.
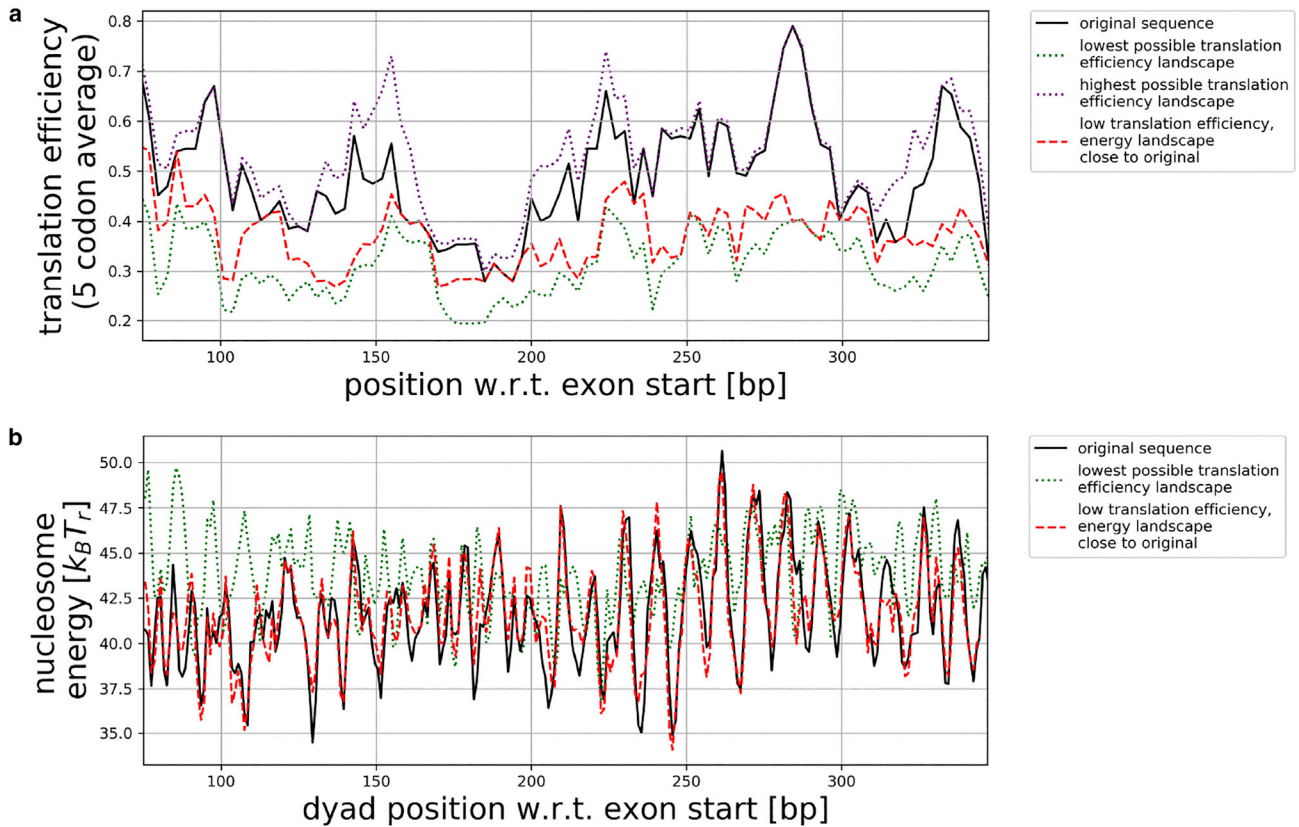
a

b

FIGURE 8   Same as Fig. 7, but focusing on low translation efficiency. To see this figure in color, go online.

Restoring all layers of information in a host organism

We next attempt to restore the translation efficiency landscape while keeping the nucleosome energy landscape in consideration. To do so, we compare ranges of five codons, the same length of DNA we study for the translation efficiency averages. To do this perfectly, one should in principle compare ranges of 147 bp, the length of a nucleosome. This, however, is impossible to achieve using our method as the graphs would consist of too many nodes. Fortunately we will see that it is not necessary to be so precise. Formally, we minimize $V_{T\&E}$, which is a combination of squared deviations of translation efficiency and nucleosome energy landscape between G and G''. We want to find a sequence G'' such that

$$V_{T\&E}(G, G'') \leq V_{T\&E}(G, X) \text{ for all } X \in A_G \quad (9)$$

with

$$V_{T\&E}(G, X) \equiv \sum_{p=2}^{N-3} c_T \left[\Delta T_{\text{host}}^{\text{source}}(G, X, p) \right]^2 + c_E [\Delta E(G, X, p)]^2. \quad (10)$$

The constants c_T and c_E can be freely chosen, depending on which quantity, translation efficiency, or nucleosome energy, one finds more important to be close to the original. The function $\Delta T_{\text{host}}^{\text{source}}(G, X, p)$ was defined by Eq. 8 and still describes the difference between the translation efficiency of sequence G in human and sequence X in yeast of five codons around codon position p. We introduced a function $\Delta E(G, X, p)$, which describes the same but for energy.

To properly define this function, it needs to reflect that we want to know the effect of the change of sequence on the *entire* nucleosome energy landscape. Therefore, we find $\Delta E(G, X, p)$ by summing over all possible positions of this 15 bp stretch on $147 + 14$ possible positions on a nucleosome. We sum over $147 + 14$ positions, since this is the number of positions where at least one of the possibly changed basepairs is contained within a nucleosome, i.e., the number of positions where the nucleosome energy could be affected by substitutions of codons. This leads to the definition:

$$\Delta E(G, X, p) \equiv \sum_{j=-7}^{147+7-1} \left| \sum_{i=-7}^{i=7-2} E_{j+i}(G_{p+i}, G_{p+i+1}, G_{p+i+2}) \right.$$

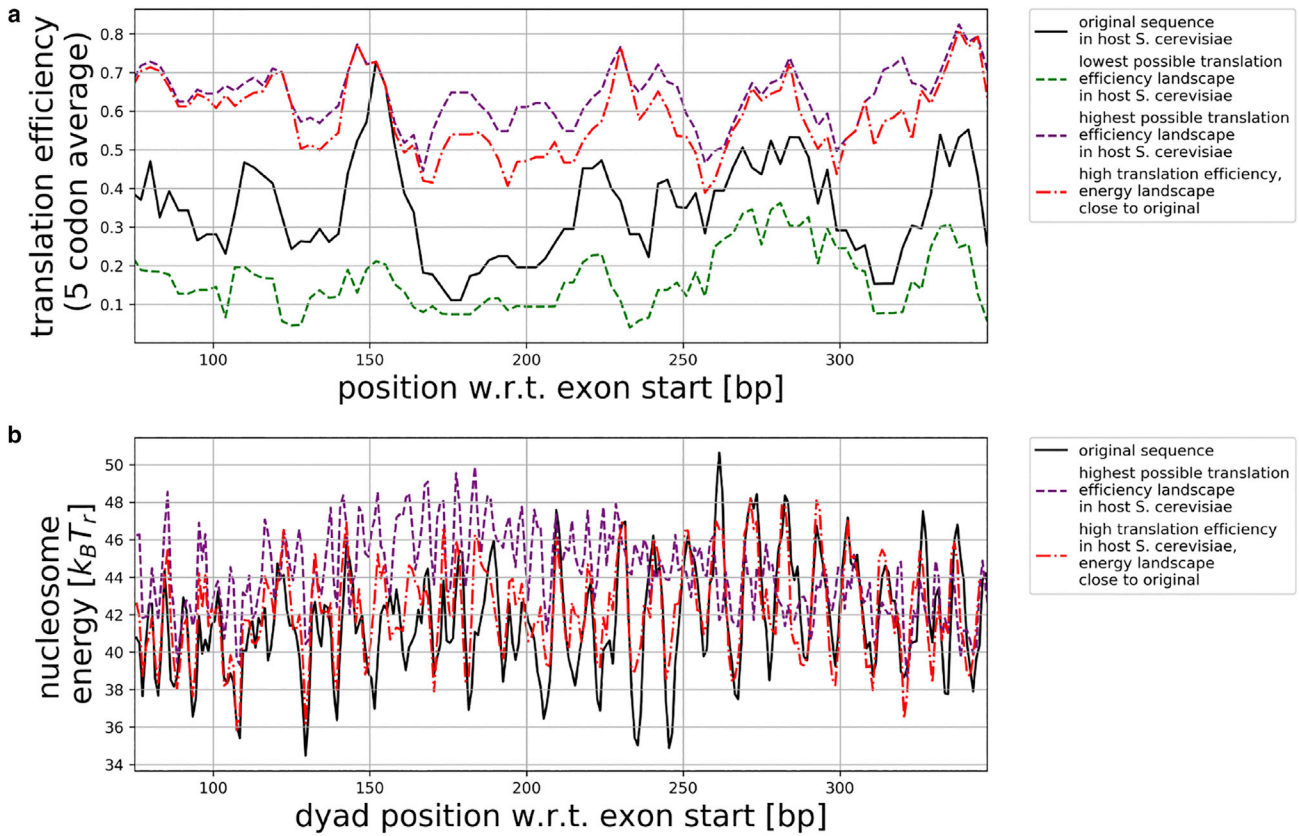$$\left. - E_{j+i}(X_{p+i}, X_{p+i+1}, X_{p+i+2}) \right|. \quad (11)$$

FIGURE 9 Same as Fig. 7, but for the human gene TNF in host organism *S. cerevisiae*. To see this figure in color, go online.

Since the nucleosome energy is invariant under a change of organism, this function too does not depend on the organisms chosen. Note that $\Delta E(G, X, p)$, like $\Delta T_{host}^{source}(G, X, p)$, is related to a total distance between the original sequence G and altered sequence X, but in this case the total distance between the nucleosome energy landscapes. This distance $D_E(G, X)$ is defined by

$$D_E(G, X) \equiv \sum_{p=2}^{N-3} \Delta E(G, X, p). \tag{12}$$

Returning to Eq. 10, we choose $c_E = 1/50000$ $[1/(k_B T_r)]$ and $c_T = 1$, which brings the quantities of efficiency and energy to the same order of magnitude while fixing the units. We introduced the squares in this equation to improve the balance between the minimization of the two quantities. It specifically aims to avoid scenarios where, e.g., ΔT is lowered significantly while an already large ΔE increases slightly, resulting in a tiny ΔT but large ΔE instead of small values of both ΔT and ΔE. The supporting material describes how to create a graph with the correct weights to obtain G''. The time complexity of "solving" the graph by using a shortest path algorithm is only linear with respect to exon size and, therefore, this method can be, in principle, used on very large exons or even on entire genomes. The

time complexity is linear because of the shape of the graph (see section 1.3 of the supporting material).

The graph (like all graphs used in this work) is more or less similar to graph $\mathscr{G}_E$ in Fig. 1. It is similar in the sense that it consists of a start node connected to a column of nodes corresponding to a certain basepair position, and any column is connected to the next column, up to the end node. In this case, the number of columns depends on the length of the exon. If the exon is very long, the graph could become too big to compute. Fortunately, the shortest path through this graph can be calculated by breaking it down in small subgraphs. These subgraphs contain two neighboring columns, through which one can calculate the shortest distances to any node in a certain column.

The result for TNF is depicted by the red dash-dotted line in Fig. 6, *a* and *b*, where it can be seen that both the nucleosome energy and the translation efficiency landscape are now close to the original.

Efficient genes with restored energy landscapes

We have studied the scenario of putting a gene in a host organism and restoring its layers of information. Now we attempt the other scenario: creating highly efficient exons while keeping the nucleosome energy landscape close to
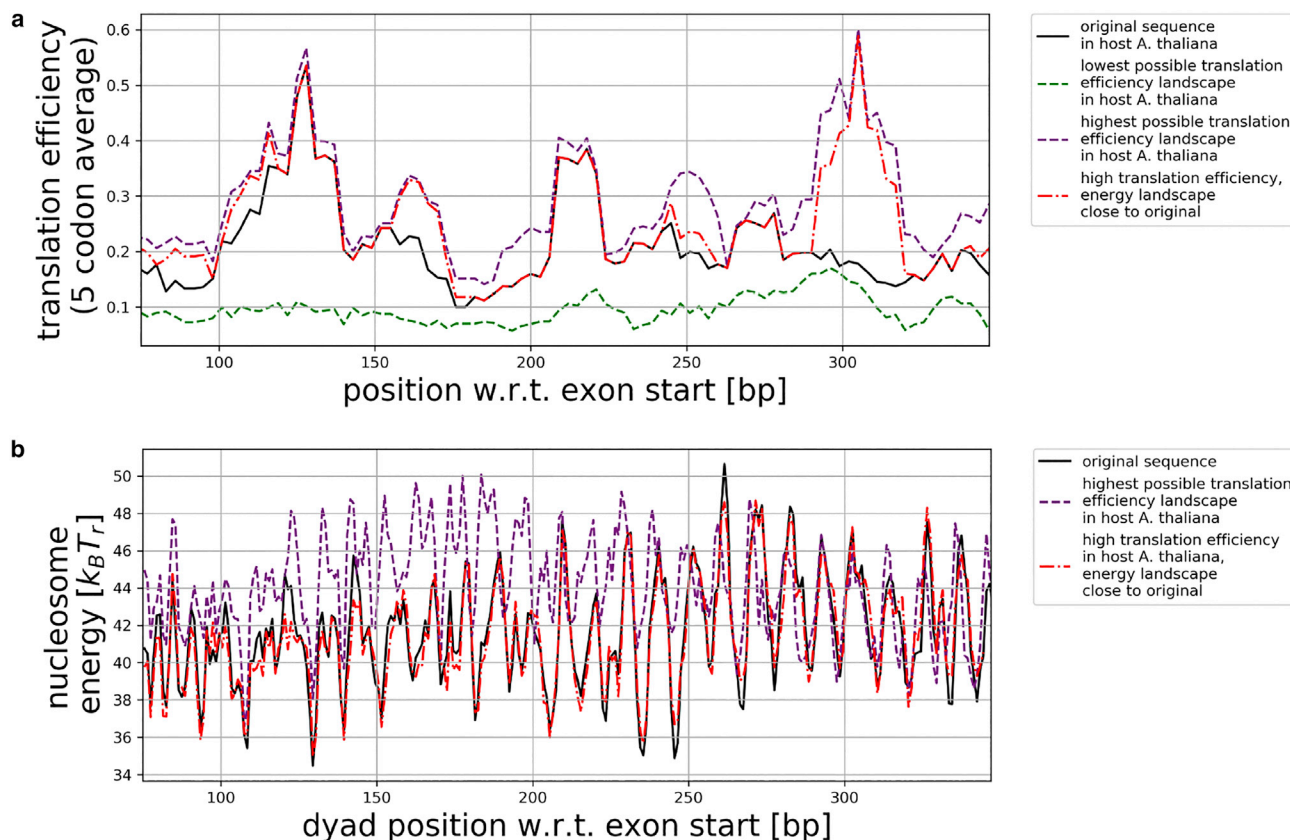
FIGURE 10   Same as Fig. 7, but for the human gene TNF in host organism *A. thaliana*. To see this figure in color, go online.

the original. The same method can be applied as in the previous case, with minor changes: $\Delta T_{\text{host}}^{\text{source}}$ in Eq. 10 should be replaced by $\Delta T_{\text{host}}^{\text{max}}$, where max corresponds to the codon sequence with the highest possible translation efficiency, not to the original efficiency (min in case of the lowest possible efficiency):

$$
V_{T\&E}^{\text{max}}(G, X) \equiv \sum_{p=2}^{N-3} c_T \left[\Delta T_{\text{host}}^{\text{max}}(G, X, p) \right]^2 \tag{13}
$$
$$
+ c_E [\Delta E(G, X, p)]^2.
$$

Again the supporting material describes how to minimize $V_{T\&E}^{\text{max}}(G, X)$ using a graph.

We demonstrate this method first using TNF in human (in this case, the host is the same as the source). The result is shown in Fig. 7. This figure is the same as Fig. 3, *a* and *b*, but with the addition of a red interrupted curve that is the result of our method. The method has restored the nucleosome energy landscape to a great extent while the translation efficiency remains close to optimal. Fig. 8 depicts the same but for a purposefully inefficient version of the exon. Using our method, we find a much lower translation efficiency and a nucleosome

landscape almost equal to the original. This method also works on TNF in yeast and *A. thaliana* (see Figs. 9 and 10). Part of the obtained sequences can be found in the supporting material.

CONCLUSION

We have presented a novel approach to study the multiplexing of genetics, mechanics, and translation efficiency. In previous work we found the highest and lowest possible nucleosome energies on top of a gene, when one can only replace codons with synonymous codons, i.e., requiring that the sequence of amino acids remains unchanged. In this work we have included the translation efficiency in our analysis, since this can be an important factor for the proper function of the final protein. One of our approaches was to add an additional restriction to the analysis: any altered sequence must have a translation efficiency landscape close to the landscape corresponding with the unaltered sequence. This restriction was applied by pruning nodes from a graph.

The second approach we used was to incorporate translation efficiency in the weights of graphs. When one puts a gene of one organism into a host organism, the translation

efficiency landscape in the host may be very different from the landscape in the original species. Using this second approach we demonstrate how to change the genetic sequence such that the host will produce a protein with a translation efficiency landscape, as well as a nucleosome energy landscape, very similar to the landscapes in the original organism. The same approach was effective in creating high translation efficiencies while keeping energy landscapes close to the original.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2022.10.011.

AUTHOR CONTRIBUTIONS

M.Z. and H.S. designed the study. B.S. and M.Z. researched an appropriate model for translation efficiency. M.Z. created the methods and performed the analyses. M.Z. and H.S. wrote the article.

ACKNOWLEDGMENTS

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Trifonov, E. N. 1989. The multiple codes of nucleotide sequences. *Bull. Math. Biol.* 51:417–432.

2. Luger, K., A. W. Mäder, …, T. J. Richmond. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature.* 389:251–260.

3. Olson, W. K., A. A. Gorin, …, V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95:11163–11168.

4. Segal, E., Y. Fondufe-Mittendorf, …, J. Widom. 2006. A genomic code for nucleosome positioning. *Nature.* 442:772–778.

5. Trifonov, E. N., and J. L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA.* 77:3816–3820.

6. Satchwell, S. C., H. R. Drew, and A. A. Travers. 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191:659–675.

7. Eslami-Mossallam, B., H. Schiessel, and J. van Noort. 2016. Nucleosome dynamics: sequence matters. *Adv. Colloid Interface Sci.* 232:101–113.

8. Kaplan, N., I. K. Moore, …, E. Segal. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature.* 458:362–366.

9. Lowary, P. T., and J. Widom. 1997. Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc. Natl. Acad. Sci. USA.* 94:1183–1188.

10. Eslami-Mossallam, B., R. D. Schram, …, H. Schiessel. 2016. Multiplexing genetic and nucleosome positioning codes: a computational approach. *PLoS One.* 11, e0156905.

11. Zuiddam, M., R. Everaers, and H. Schiessel. 2017. Physics behind the mechanical nucleosome positioning code. *Phys. Rev. E.* 96, 052412.

12. Zuiddam, M., and H. Schiessel. 2019. Shortest paths through synonymous genomes. *Phys. Rev. E.* 99, 012422.

13. Tillo, D., and T. R. Hughes. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinf.* 10:442.

14. Struhl, K., and E. Segal. 2013. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* 20:267–273.

15. Drillon, G., B. Audit, …, A. Arneodo. 2016. Evidence of selection for an accessible nucleosomal array in human. *BMC Genom.* 17:526.

16. Tompitak, M., C. Vaillant, and H. Schiessel. 2017. Genomes of multicellular organisms have evolved to attract nucleosomes to promoter regions. *Biophys. J.* 112:505–511.

17. Neipel, J., G. Brandani, and H. Schiessel. 2020. Translational nucleosome positioning: a computational study. *Phys. Rev. E.* 101, 022405.

18. Vavouri, T., and B. Lehner. 2011. Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. *PLoS Genet.* 7, e1002036.

19. Studitsky, V. M., D. J. Clark, and G. Felsenfeld. 1994. A histone octamer can step around a transcribing polymerase without leaving the template. *Cell.* 76:371–382.

20. Bednar, J., V. M. Studitsky, …, C. L. Woodcock. 1999. The nature of the nucleosomal barrier to transcription: direct observation of paused intermediates by electron cryomicroscopy. *Mol. Cell.* 4:377–386.

21. Kujirai, T., H. Ehara, …, H. Kurumizaka. 2018. Structural basis of the nucleosome transition during RNA polymerase II passage. *Science.* 362:595–598.

22. Polach, K. J., and J. Widom. 1995. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.* 254:130–149.

23. Hodges, C., L. Bintu, …, C. Bustamante. 2009. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science.* 325:626–628.

24. Culkin, J., L. de Bruin, …, H. Schiessel. 2017. The role of DNA sequence in nucleosome breathing. *Eur. Phys. J. E Soft Matter.* 40:106.

25. Lowary, P. T., and J. Widom. 1998. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* 276:19–42.

26. Mauney, A. W., J. M. Tokuda, …, L. Pollack. 2018. Local DNA sequence controls asymmetry of DNA unwrapping from nucleosome core particles. *Biophys. J.* 115:773–781.

27. van Deelen, K., H. Schiessel, and L. de Bruin. 2020. Ensembles of breathing nucleosomes: a computational study. *Biophys. J.* 118:2297–2308.

28. Bondarenko, V. A., L. M. Steele, …, V. M. Studitsky. 2006. Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. *Mol. Cell.* 24:469–479.

29. Schwartz, S., E. Meshorer, and G. Ast. 2009. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* 16:990–995.

30. Depken, M., J. M. R. Parrondo, and S. W. Grill. 2013. Intermittent transcription dynamics for the rapid production of long transcripts of high fidelity. *Cell Rep.* 5:521–530.

31. Nogues, G., S. Kadener, …, A. R. Kornblihtt. 2002. Transcriptional activators differ in their abilities to control alternative splicing. *J. Biol. Chem.* 277:43110–43114.

32. de la Mata, M., C. R. Alonso, …, A. R. Kornblihtt. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Mol. Cell.* 12:525–532.

33. Li, Y., C. Li, …, C.-Y. Li. 2018. Human exonization through differential nucleosome occupancy. *Proc. Natl. Acad. Sci. USA.* 115:8817–8822.

34. Meersseman, G., S. Pennings, and E. M. Bradbury. 1992. Mobile nucleosomes–a general behavior. *EMBO J.* 11:2951–2959.

35. Kulić, I. M., and H. Schiessel. 2003. Chromatin dynamics: nucleosomes go mobile through twist defects. *Phys. Rev. Lett.* 91, 148103.

36. Mohammad-Rafiee, F., I. M. Kulić, and H. Schiessel. 2004. Theory of nucleosome corkscrew sliding in the presence of synthetic DNA ligands. *J. Mol. Biol.* 344:47–58.

37. Brandani, G. B., T. Niina, …, S. Takada. 2018. DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations. *Nucleic Acids Res.* 46:2788–2801.

38. Schiessel, H., J. Widom, …, W. M. Gelbart. 2001. Polymer reptation and nucleosome repositioning. *Phys. Rev. Lett.* 86:4414–4417.

39. Kulić, I. M., and H. Schiessel. 2003. Nucleosome repositioning via loop formation. *Biophys. J.* 84:3197–3211.

40. Lequieu, J., D. C. Schwartz, and J. J. de Pablo. 2017. In silico evidence for sequence-dependent nucleosome sliding. *Proc. Natl. Acad. Sci. USA.* 114:E9197–E9205.

41. Niina, T., G. B. Brandani, …, S. Takada. 2017. Sequence-dependent nucleosome sliding in rotation-coupled and uncoupled modes revealed by molecular simulations. *PLoS Comput. Biol.* 13, e1005880.

42. Guo, A. Z., J. Lequieu, and J. J. de Pablo. 2019. Extracting collective motions underlying nucleosome dynamics via nonlinear manifold learning. *J. Chem. Phys.* 150, 054902.

43. Rudnizky, S., H. Khamis, …, A. Kaplan. 2019. The base pair-scale diffusion of nucleosomes modulates binding of transcription factors. *Proc. Natl. Acad. Sci. USA.* 116:12161–12166.

44. Winger, J., I. M. Nodelman, …, G. D. Bowman. 2018. A twist defect mechanism for ATP-dependent translocation of nucleosomal DNA. *Elife.* 7, e34100.

45. Li, M., X. Xia, …, Z. Chen. 2019. Mechanism of DNA translocation underlying chromatin remodelling by Snf2. *Nature.* 567:409–413.

46. Sabantsev, A., R. F. Levendosky, …, S. Deindl. 2019. Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling. *Nat. Commun.* 10:1720.

47. Brandani, G. B., and S. Takada. 2018. Chromatin remodelers couple inchworm motion with twist-defect formation to slide nucleosomal DNA. *PLoS Comput. Biol.* 14, e1006512.

48. Segal, E., and J. Widom. 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* 19:65–71.

49. Schiessel, H., and R. Blossey. 2020. Pioneer transcription factors in chromatin remodeling: the kinetic proofreading view. *Phys. Rev. E.* 101, 040401.

50. Dong, H., L. Nilsson, and C. G. Kurland. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J. Mol. Biol.* 260:649–663.

51. Gingold, H., D. Tehler, …, Y. Pilpel. 2014. A dual program for translation regulation in cellular proliferation and differentiation. *Cell.* 158:1281–1292.

52. Stein, K. C., and J. Frydman. 2019. The stop-and-go traffic regulating protein biogenesis: how translation kinetics controls proteostasis. *J. Biol. Chem.* 294:2076–2084.

53. Collart, M. A., and B. Weiss. 2020. Ribosome pausing, a dangerous necessity for co-translational events. *Nucleic Acids Res.* 48:1043–1055.

54. López, D., and F. Pazos. 2015. Protein functional features are reflected in the patterns of mRNA translation speed. *BMC Genom.* 16:513.

55. Pechmann, S., and J. Frydman. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20:237–243.

56. Zhou, M., J. Guo, …, Y. Liu. 2013. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature.* 495:111–115.

57. O'Brien, E. P., P. Ciryam, …, C. M. Dobson. 2014. Understanding the influence of codon translation rates on cotranslational protein folding. *Acc. Chem. Res.* 47:1536–1544.

58. Liutkute, M., E. Samatova, and M. V. Rodnina. 2020. Cotranslational folding of proteins on the ribosome. *Biomolecules.* 10:97.

59. Tompitak, M., G. T. Barkema, and H. Schiessel. 2017. Benchmarking and refining probability-based models for nucleosome-DNA interaction. *BMC Bioinf.* 18:157.

60. Frank, J., and R. L. Gonzalez, Jr. 2010. Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu. Rev. Biochem.* 79:381–412.

61. Wohlgemuth, I., C. Pohl, …, M. V. Rodnina. 2011. Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366:2979–2986.

62. dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32:5036–5044.

63. Sabi, R., and T. Tuller. 2014. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res.* 21:511–526.

64. Yoon, J., Y. J. Chung, and M. Lee. 2018. STADIUM: species-specific tRNA adaptive Index compendium. *Genomics Inform.* 16:e28.

65. Lodish, H., A. Berk, …, M. P. Scott. 2013. Molecular Cell Biology, Seventh edition. W. H. Freeman and Company.

66. Chu, W. M. 2013. Tumor necrosis factor. *Cancer Lett.* 328:222–225.

67. Dolgin, E. 2017. The most popular genes in the human genome. *Nature.* 551:427–431.